Text Normalization Challenge - English

C. Arenas, M. Combalia, D. Moreno, G. Lahuerta

Universitat Politècnica de Catalunya

21st November

Overview

1. Introduction
   - Motivation
   - Goal
2. Environment
   - Dataset
3. Environment
   - Dataset
   - Proposed framework
4. Set Up
5. Current results

Introduction

Motivation

Main speech and language applications:

- Text to Speech (TTS)
- Automatic Speech Recognition (ASR)

Normalization is needed to adapt some written forms in order to be understood by the application

Introduction

Motivation

Main speech and language applications:

- Text to Speech (TTS)
- Automatic Speech Recognition (ASR)

Normalization is needed to adapt some written forms in order to be understood by the application

Some Examples

In the following table we can see some examples of written forms that do not match the spoken form:

Written form	Spokem form
150lb	One hundred and fifty pounds
$3.16	Three dollars, sixteen cents
12:47	Twelve forty-seven

Introduction

Motivation

Normalization is needed for the good work of the voice assistants.

- Google Assistant (Google)
- Cortana (Windows)
- Siri (Apple)
- Alexa (Amazon)

Introduction

Motivation

This challenge belongs to a real kaggle competition

- 289 participants registered
- 270 teams formed

Rewards:

- 1st place: 12.000€
- 2nd place: 8.000€
- 3rd place: 5.000€

Introduction

Goal

The main goal of this project is to develop a system that:

- Adapt an English-German translation NN to the challenge dataset
- Normalizes the written expressions into readable expressions

Environment

Dataset I

Example (Dataset)

```
seqID, wordID, wordType, src, dst
0,0,”PLAIN”,”Brillantaisia”,”Brillantaisia”
0,1,”PLAIN”,”is”,”is”
0,2,”PLAIN”,”a”,”a”
0,3,”PLAIN”,”genus”,”genus”
0,4,”PLAIN”,”of”,”of”
0,5,”PLAIN”,”plant”,”plant”
0,6,”PLAIN”,”in”,”in”
0,7,”PLAIN”,”family”,”family”
0,8,”PLAIN”,”Acanthaceae”,”Acanthaceae”
```

Environment

Dataset II

Example (Dataset)

Brillantaisia is a genus of plant in family Acanthaceae .

University of California publications in linguistics (vol 35) .

Retrieved April 10, 2013 .

However , Kaede's oiled hair accidentally catches fire .

Tropical Depression Two was upgraded into a tropical storm in post season analysis , and as a result has no name .

Retrieved on November 4, 2014 .

Retrieved 7 December 2015 .

Francisca Josefa had three siblings whose names were Catalina and Pedro Antonio Diego

Henry moved in with Young at Carrigoona Cottage in 1929 , building a studio there .

Retrieved 3 February 2016 .

Environment

Proposed framework

- **Why?**
  PyTorch is a reliable language

Environment

Proposed framework

PYTORCH

Deep Learning with PyTorch

- **Why?**
  PyTorch is a reliable language
- **Advantages**
  More versatile than keras
  Easier to use than Tensorflow

Environment

Proposed framework

- **Why?**
  PyTorch is a reliable language
- **Advantages**
  More versatile than keras
  Easier to use than Tensorflow
- **Accessible material**
  We already have some code focused on translate from English to German

Set Up

Problem Analysis

Our problem, in reality, is a Machine Translation problem. From 'correct' English to a spoken one.

Set Up

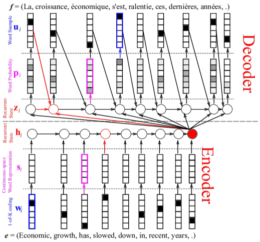
Problem Analysis

Our problem, in reality, is a Machine Translation problem. From 'correct' English to a spoken one.

Database Examples

'Correct' English	Common Spoken English
Herry was born in Banda Aceh , Aceh from a Sundanese descendant.	Herry was born in Banda a c e h , a c e h from a Sundanese descendant .
1987 ALCS Game 2 - Detroit Tigers vs	nineteen eighty seven a l c s Game two - Detroit Tigers versus

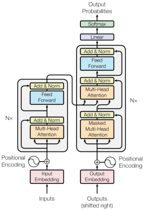
Set Up


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Sample u_i
Word Probability p_i
Recurrent State z_i
Decoder
Recurrent State h_i
Continuous-space Word Representation s_i
1-of-K coding w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Lazy Machine Translation State of the Art

1. Statistical Machine Translation
2. Neural Machine Translation → Recurrent Neural Networks
3. Attention

Set Up



"Attention is all you need", The Transformer

- No RNN, just Attention mechanism
- Google Design
- Simple
- Fast
- State of the Art results this summer

Current results

Right now, our NN is training...

```
-Epoch 4:-
- (Training)
ppl:  1.58991,
    accuracy: 93.459 %,
    elapse: 153.527 min
- (Validation)
ppl:  1.49141,
    accuracy: 94.243 %,
    elapse: 11.279 min
```

Bibliography

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
CoRR, abs/1706.03762, 2017.

Questions?